Estimating the Number Of Unseen Studies in a Meta-Analysis

L. E. Eberly* and George Casella†
Biometrics Unit
Cornell University
Ithaca, NY 14853
BUM # 1308-MA MB

February 24, 1997

Abstract

A parameter estimate from a meta-analysis is biased when the experiments to be combined are a non-random sample from the population of all experiments done on the hypothesis of interest. In particular, publication bias occurs when studies with significant results are more likely to be published than studies with non-significant results. We develop a model for the distribution of the total number of studies carried out, both published and unpublished, dependent on the probability of publication. We assume a selection model where all studies significant at level α are published, while non-significant studies are published with probability ρ. Using Metropolis simulation and Gibbs sampling techniques, we generate random samples from the distribution of the total number of studies and study how it changes as ρ varies. An application on lead exposure and IQ level in children is presented and the results interpreted. Comparisons are made with Rosenthal's fail-safe estimators.

Keywords: Bayesian modeling, file-drawer problem, Gibbs sampling, publication bias, Rosenthal's fail-safe numbers

*Supported by National Institute of Environmental Health Sciences Training Grant EHS-5-T32-ES07261-03 and National Science Foundation Grant DMS-9305547.

†Supported by National Science Foundation Grant DMS-9305547.

1 Introduction

Meta-analysis, a method of combining results from different experiments on the same hypothesis, has gained wide recognition in both the statistical and the scientific worlds in the past twenty years. The primary concern when carrying out a meta-analysis is the validity and reliability of the resulting overall conclusion, *e.g.*, the test of an effect estimate. The most common method of finding experimental results to include in a meta-analysis on a specific hypothesis is through literature searches in relevant journals. However, journals can be unrepresentative for a number of reasons. Often studies with statistically non-significant results are underrepresented in the literature. For example, a scientist may not submit the results of a study that does not show some statistically significant result, or a journal editor may not accept those results, either one feeling that a result of "no difference" would be of little importance to the scientific community. Thus, any sample of studies from the published literature is typically non-random. When a meta-analysis of these studies is then done, an overall effect estimate could be biased towards a higher level of significance (Hedges, 1992). According to Bayarri and DeGroot (1986), selection bias is the distortion in an effect estimate resulting when a non-random sample is drawn from the population of interest. We concern ourselves with publication bias in particular, the selection bias resulting when studies statistically significant at some level α are more likely to be published than non-significant studies.

Easterbrook, *et al.* (1991) carried out a retrospective study of 285 analyzed research projects which had been approved by the Central Oxford Research Ethics Committee between 1984 and 1987 in order to show that publication bias does in fact exist in the medical literature. Using logistic regression and adjusting for relevant covariates, they found that projects with statistically significant results (defined to have a p-value < 0.05) were more likely to have been published and/or presented than those with non-significant results (odds ratio=3.56, 95% C.I.=(1.82,6.99)).

In addition, they noted that 43 of the 78 unpublished projects had obtained null results. Only 8 of those 43 were written up and subsequently rejected, while 26 were never written up *because* they showed null results. Thus, it appears that publication bias here is primarily due to lack of submission of null results, not because of editorial rejection of submitted null results. Dickersin, *et al.* (1992) carried out a similar study using research projects that appeared on the institutional review board logs for the Johns Hopkins Health Institutions, including the School of Medicine, Hospital, Kennedy Institute, School of Nursing, and the Frances Scott Key Medical Center, and the School of Hygiene and Public Health. Using logistic regression and adjusting for covariates, they found similar results, even the conclusion that the problem lies with authors, not editors.

A variety of methods for dealing with publication bias have been proposed. Rosenthal (1979) began with the fail-safe number, which calculates the number of unseen studies averaging null results needed to bring a meta-analytic result to some pre-specified level of significance. White (1982) and Glass, *et al.* (1981) suggest obtaining results for studies which were not published (through surveys of colleagues, for example, or national registries of studies) and comparing those results to the published results. Light and Pillemer (1984) describe a method to detect publication bias using a "funnel graph" of sample size *vs.* effect estimate. In the presence of publication bias, and assuming effect size is unrelated to sample size, the graph should be missing the lower left-hand corner of the pyramid. Berlin, Begg, and Louis (1989) introduce a method to quantify the information in a funnel graph by using a model relating bias to sample size under the same assumption. Results indicated that small trials are more prone to publication bias and that the bias may be substantial, especially when the trial was based on a non-randomized design. A more recent extension of the funnel graph idea (in Begg and Mazumdar, 1994) suggests calculating and then testing a rank correlation between effect estimates and their variances. A positive correlation

would indicate that negative studies are less likely to be published.

The funnel graph and correlation approaches have the advantage of being based on assumptions which are distribution-free. Hedges (1984) meanwhile pursued truncated sampling models, where it was assumed that statistically non-significant results (at α-level=0.05) do not get published. He found that the bias can depend on a study's sample size and effect size, and can be substantial for either small samples or small effects. Bayarri and DeGroot (1986, 1991) explore the behavior of published results using an indicator function of statistical significance to weight the model's likelihood, and show that significant overall results obtained from published data actually can be strongly supportive of the null hypothesis. Iyengar and Greenhouse (1988) modify Bayarri and DeGroot's methods slightly by not restricting the selection to this "publish if and only if significant" situation. They incorporate a family of weight functions into the model's likelihood, using the conditional probability of reporting a study given the data as the weight, where this probability varied across studies. Hedges (1992) and Dear and Begg (1992) take the same approach, but modify these weight functions slightly, while Cleary (1996) computes estimates of the parameter of interest as a function of the selection parameter. Frongillo (1991) takes a Bayesian approach and uses two-stage hierarchical models to model variability both within and between studies.

Historically, then, there have been three general methods of dealing with publication bias: truncated sampling models, invariant sampling, and source augmentation. Truncated sampling models assume that no non-significant studies are published, and then, usually through simulations, determine the bias in the effect estimate that comes about due to the publication process. Recently this has been extended to include less strict selection processes. Invariant sampling methods limit the meta-analysis to a subset of studies which come from a sampling frame independent of the publication process (*e.g.*, registries of studies); extensive registries of studies, though, do not as yet exist in

most fields of research. Source augmentation speculates on the number of missing (unpublished) studies and may then adjust effect estimates accordingly (Begg and Berlin, 1988). Of the three methods, truncated sampling and invariant sampling often assume that the researcher has access to each study's effect estimates and perhaps sample variances. Reality forces us to acknowledge, though, that often we cannot acquire the original data from a study, sometimes not even the effect size estimates. Especially with older studies, it is likely that only p-values or t-values can be gleaned from the publication itself. This renders the use of many of the above methods impossible. On the contrary, source augmentation methods that have been developed so far (as well as the one we will explore) do not require more than p- or t-values. In spite of this advantage, we believe source augmentation should, whenever possible, be carried out *in addition to* effect size estimation. Both are important aspects of a meta-analysis.

In this paper, we model the distribution of the total number of studies carried out, both seen and unseen, dependent on the probability of publication. This method still necessitates estimating a selection probability, but the distribution can then be calculated for a range of probability values, leading at least to a somewhat more detailed picture. Section 2 of this paper covers the derivation of the model and the assumptions associated with it, including the sampling methods used. Section 3 explains the results from simulations based on the model. Section 4 presents an application of the results to a meta-analysis on studies of lead exposure and IQ levels in children, and makes comparisons to the standard source augmentation method, Rosenthal's fail-safe number. Section 5 presents our conclusions regarding the uses and limitations of this theory, and directions for further research.

2 The Approach

2.1 The Model

Throughout this paper, we assume that some of the assumptions necessary to conduct a meta-analysis hold:

i. Each of the observed studies tests the same hypothesis.

ii. The observed studies are independent.

The following is a usual assumption of meta-analysis that we presume does *not* hold:

iii. The observed studies are a random sample from the population of all studies that have been carried out on this hypothesis.

Most researchers agree that *some* form of selection bias, particularly publication bias, is present in any field, which invalidates assumption (iii.). The probability of publication, call it Q, quite likely varies widely from field to field, from journal to journal, and from year to year. We will impose a prior distribution on Q in order to account for this variability using a Beta distribution:

$$\pi_Q(q|a,b) = \frac{1}{B(a,b)} q^{a-1}(1-q)^{b-1}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, $\Gamma(x) = \int_0^\infty t^{x-1}\exp^{-t} dt$, $0 \leq q \leq 1$, $a > 0$, and $b > 0$. (Throughout this paper, we will use the symbol π to denote probability mass or density functions.) The Beta distribution is very flexible, and by its parameters can vary from a bathtub shape through a uniform to a bell-shaped distribution. Assuming publication bias is present, Q must be dependent on the probability of achieving a statistically significant result, call it R. We can write:

$$
\begin{aligned}
Q &= P[\text{publication}|\text{significant}]\, P[\text{significant}] \\
&\quad + P[\text{publication}|\text{non-significant}]\, P[\text{non-significant}] \\
&= R + \rho(1-R). \qquad (1)
\end{aligned}
$$

This structure dictates that all significant studies are published, and that the proportion ρ of the non-significant studies are published. ρ is a selection parameter; we will treat it as a known value. If $\rho = 1$, then every study will be published with probability 1; if $\rho = 0$, then a study will be published with the probability R that it is significant.

When conducting a meta-analysis, one reviews the available literature and finds all published studies that test the hypothesis of interest. If k such studies are found, there are still an unknown number, call it $N - k$, of studies that have actually been done, but were *not* published. We can thus model this using a negative binomial distribution: how many studies does it take until we see k successes? Thus the number of studies total is the random variable, N:

$$
\pi_N(n|q,k) = \binom{n-1}{k-1} q^k (1-q)^{n-k}, \qquad (2)
$$

where $n = k, k+1, \ldots, k \in \mathbf{W}$, and $0 \leq q \leq 1$. The distribution of N that we have here is conditional on an unknown value, namely q. What we are ultimately interested in, however, is the marginal distribution of N which is no longer dependent on the value of q.

It is easy to find this marginal through the calculation $\pi_N(n|k) = \int \pi_N(n|q,k)\pi_Q(q)\, dq$, but the only observed data that this incorporates is the number of published studies, k. We are ignoring important information relevant to publication bias if we don't take into account the number of *significant* published studies. As we shall see in Section 2.2, incorporating this information into our model makes it much more difficult to find $\pi_N(n|k)$. This can be easily obtained through a Gibbs sampling procedure (as will be explained in Section 2.3), but the procedure requires our model's

full conditional specification:

$$\pi_N(n|q,\theta,\text{data}) \quad \text{and} \quad \pi_Q(q|n,\theta,\text{data}), \tag{3}$$

These are the conditional distributions of each unknown parameter of interest, where θ denotes (ρ, a, b), the nuisance parameters.

2.2 Derivation of the Full Conditional Specification

We will first derive the conditional distribution of Q given $N = n$. Given $\pi_N(n|q,k)$ and $\pi_Q(q|a,b)$, we can derive $\pi_Q(q|n,k,a,b)$. We then have:

$$\pi_Q(q|n,k,a,b) = \frac{1}{B\left(k+a, n-k+b\right)} q^{k+a-1}\left(1-q\right)^{n-k+b-1},$$

where $0 \leq q \leq 1$. As stated above, we have no data in this model yet except for k. Consider the formulation of Q given in Equation 1. Given $R = r$ and a pre-specified level of significance α, any study will be significant with probability r and non-significant with probability $1 - r$. Assuming studies are independent (which is not too unreasonable), every study we consider, observed or unobserved, is the realization of a Bernoulli(r) random variable. Since larger studies will have more power, and hence are actually more likely to achieve statistical significance, we need to assume that the studies are of approximately the same size; then r will be constant across studies. The k observed studies in particular are thus k independent Bernoulli trials, of which a certain number will be "successes," where a success means statistical significance. This leads us to a Binomial(k, r) random variable, call it Z, which counts the number of *significant* studies within the *observed* studies:

$$\pi_Z(z|k,r) = \binom{k}{z} r^z \left(1-r\right)^{k-z},$$

where $z = 0, 1, \ldots, k$ and $0 \leq r \leq 1$. (r does depend on the chosen significance level α and could be denoted r_α.) The usual estimate of r is $\hat{r} = z/k$, the maximum likelihood estimator (MLE). To then get estimates of Q in a given situation, use from Equation 1: $\hat{q} = \hat{r} + \rho(1 - \hat{r}) = (1-\rho)z/k + \rho$, the MLE of Q. We can now calculate the distribution of the MLE $\hat{Q}$, dependent on the values of q, k, r, and ρ:

$$\begin{aligned} \pi_{\hat{Q}}(\hat{q}|q,k,r,\rho) &= P\left[\,(1-\rho)\,Z/k + \rho = \hat{q} \mid k,r,\rho\,\right] \\ &= P\left[Z = \tfrac{k(\hat{q}-\rho)}{1-\rho} | k,r,\rho\right] \\ &= \binom{k}{\frac{k(\hat{q}-\rho)}{1-\rho}} r^{\frac{k(\hat{q}-\rho)}{1-\rho}} (1-r)^{k-\frac{k(\hat{q}-\rho)}{1-\rho}} \end{aligned}$$

where $\rho \leq \hat{q} \leq 1$ and $0 \leq \rho \leq 1$. Although it is not explicitly part of the equation, this density *is* dependent on q, through r and ρ by Equation 1. Note also that the introduction of the dependence on ρ leads the range of $\hat{q}$ to be bounded below by ρ.

Now that we have derived the distribution of the observed data $\hat{q}$, we should incorporate that information into our full conditional specification (Equation 3). Again using probability calculus:

$$\pi_Q(q|n,k,\rho,a,b,\hat{q}) = \frac{(q-\rho)^z\,(1-q)^{n+b-z-1}\,(q)^{k+a-1}}{\int_\rho^1 (q-\rho)^z\,(1-q)^{n+b-z-1}\,(q)^{k+a-1}\,dq}, \tag{4}$$

where $\rho \leq q \leq 1$ and $0 \leq \rho \leq 1$. We now have $\pi_Q(q|n,k,\rho,a,b,\hat{q})$ and $\pi_N(n|q,k)$. It appears as if the conditional distribution that we have for N is not the distribution needed for the full conditional specification, but note that given q, the values of ρ, a, b, and $\hat{q}$ are irrelevant. In other words, we assume that N is conditionally independent of these values. Thus, $\pi_N(n|q,k,\rho,a,b,\hat{q}) = \pi_N(n|q,k)$ and we are ready to implement the Gibbs sampler. To simplify notation, and since they are assumed known, we will suppress the dependence on ρ, a, and b from now on.

2.3 Sampling Techniques

The Gibbs sampler is an iterative Monte Carlo Markov chain simulation technique introduced by Geman and Geman (1984) and further developed by Tanner and Wong (1987) and Gelfand and Smith (1990). Very generally speaking, the purpose of the Gibbs sampler is to replace a difficult calculation (here, that of $\pi_N(n|k)$) with a sequence of easier calculations (from $\pi_N(n|q,k)$). The algorithm alternately generates values from our two distributions in Equation 3 as follows:

[0.] Choose an arbitrary starting value $q_0 \in [0,1]$.

[1.] For $i = 1, \ldots, t$, generate: n_i from $\pi_N(n|q_{i-1}, k)$

q_i from $\pi_Q(q|n_i, k, \hat{q})$.

The values of n_i and the values of q_i over the iterations form two Markov chains, $n_1, n_2, \ldots, n_t$ and $q_1, q_2, \ldots, q_t$, with transition kernels:

$$\begin{aligned} K_1(n, n') &= \int \pi_N(n|q,k)\pi_Q(q|n', k, \hat{q})\, dq \\ K_2(q, q') &= \int \pi_Q(q|n,k,\hat{q})\pi_N(n|q', k)\, dn. \end{aligned}$$

Under regularity conditions described in Geman and Geman (1984) and Tanner and Wong (1987), among many others, we have the following asymptotic results:

$$\begin{aligned} n_t &\xrightarrow{\mathcal{D}} N \quad \sim \pi_N(n|k) \quad \text{as} \quad t \to \infty \\ q_t &\xrightarrow{\mathcal{D}} Q \quad \sim \pi_Q(q|k,\hat{q}) \quad \text{as} \quad t \to \infty, \end{aligned} \tag{5}$$

independent of the starting value q_0. Recall that our goal is to examine the distribution of N, the total number of studies carried out. This asymptotic result tells us that by generating a large enough sample $n_1, n_2, \ldots, n_t$, we can determine *any* characteristic of $\pi_N(n|k)$ to *any* degree of precision.

Before we can proceed with this algorithm, however, notice that we cannot directly generate values from one of our conditional distributions, that of Q. Due to the integral in the denominator, we also cannot find a good, well-behaving approximate distribution that has a calculable (finite) maximum in order to use rejection sampling. The Metropolis method (Metropolis, *et al.*, 1953) seems to be the best option; this algorithm generates a value for Q from a "candidate" distribution, and accepts that value if it is "close enough" that it could have come from the target distribution in Equation 4. We begin with the following candidate distribution:

$$\pi_{\Psi}(\psi|n,k,\hat{q}) = \frac{1}{B\left(z+\ell+1, n+b-z\right)}\psi^{z+\ell}\left(1-\psi\right)^{n+b-z-1},$$

where $0 \leq \psi \leq 1$, and which matches the power of the $1-\psi$ term above with the $1-q$ term in Equation 4. This is a particular Beta distribution, from which it will be easy to generate samples. We leave ℓ to be determined later; this value can be fine-tuned (for various values of a and b, for example) in order to even more closely approach the target distribution of Q. To correctly simulate Q over its range from ρ to 1, though, we must transform with $\Phi = (1-\rho)\Psi + \rho$, which has the desired range and the following probability density function:

$$\pi_{\Phi}(\phi|n,k,\hat{q}) \;=\; \frac{1}{B\left(z+\ell+1, n+b-z\right)}\left(\frac{1}{1-\rho}\right)^{n+b+\ell}(\phi-\rho)^{z+\ell}\left(1-\phi\right)^{n+b-z-1}. \qquad (6)$$

According to the Metropolis algorithm, we will generate a sample $q_1, q_2, \ldots, q_t$ from the desired distribution in Equation 4 by applying a decision rule as follows: given a random Uniform$(0,1)$ number, u_i, and a random ϕ obtained as $\phi = (1-\rho)\psi + \rho$ from a random ψ:

$$u_i < \min\left\{1, \tfrac{h(\phi)}{h(q_{i-1})}\tfrac{g^*(q_{i-1})}{g^*(\phi)}\right\} \quad \Rightarrow \quad q_i = \phi$$

$$u_i > \min\left\{1, \tfrac{h(\phi)}{h(q_{i-1})}\tfrac{g^*(q_{i-1})}{g^*(\phi)}\right\} \quad \Rightarrow \quad q_i = q_{i-1}.$$

In our model,

$$h\left(q\right) = \;(q-\rho)^{z}\left(1-q\right)^{n+b-z-1}\left(q\right)^{k+a-1}$$

$$g^*\left(q\right) = \;\tfrac{1}{B(z+\ell+1,n+b-z)}\left(\tfrac{1}{1-\rho}\right)^{n+b+\ell}(q-\rho)^{z+\ell}\left(1-q\right)^{n+b-z-1}$$

from Equations 4 and 6 respectively. Simplifying, we have:

$$u_i < \min\left\{1, \left(\frac{\phi}{q_{i-1}}\right)^{k+a-1}\left(\frac{q_{i-1}-\rho}{\phi-\rho}\right)^{\ell}\right\} \quad \Rightarrow \quad q_i = \phi. \tag{7}$$

Otherwise, the Metropolis sequence does not move and $q_i = q_{i-1}$. As $i \to \infty$, then, the distribution of q_i converges to the desired target distribution. See Metropolis, *et. al* (1953) for more details.

The Gibbs technique is most easily applied here simultaneously with the Metropolis sampling as an iterative algorithm. First generate q_0; subsequently, alternate between generating from the conditional distribution of N (Equation 2) and from the *candidate* conditional distribution of Q (Equation 6). More precisely, for $i = 1, 2, \ldots, t$:

[0.] Generate $q_0 \sim Uniform(0, 1)$.

[1.] Generate: n_i from $\pi_N(n|q_{i-1}, k)$

$q_i^* = \phi$ from $\pi_\Phi(\phi|n_i, k, \hat{q})$.

[2.] Test q_i^* using the Metropolis criterion specified in Equation 7. Denote an "acceptable" value of q_i^* by q_i; otherwise, let $q_i = q_{i-1}$.

[3.] Return to [1.].

This combined algorithm also produces two Markov chains, $n_1, n_2, \ldots, n_t$ and $q_1, q_2, \ldots, q_t$, each of which converges in distribution to the desired marginal, as in Equation 5.

3 Results

3.1 Simulation Description

The simulation algorithm described in Section 2.3 is run using the GAUSS System (Version 3.01) to produce 10,000 generated numbers of each of N and Q in total, using 10 cycles of 1000 generations each. The number of published studies k is set first at 5 and then at 20, in order to see the effect of a small *vs.* a large meta-analysis situation. The prior parameters on Q (a and b) are each taken to be 5, since this gives the Beta distribution a symmetric bell-shape with somewhat thick tails. A variety of values for $\hat{Q}$ and ρ are chosen in order to see how the results varied. The value of $\hat{Q}$ is taken to be 1/10, 1/2, and 9/10 to cover as wide a range as possible. All valid values for ρ (as restricted by Equation 1) are used when $k = 5$. However, when $k = 20$, simulations are run for only 11 of the possible 33 values for ρ, approximately evenly spaced. In total, twenty sets of values are studied, nine for $k = 5$ and eleven for $k = 20$ (see Table 1).

The parameter ℓ is (somewhat arbitrarily) set at $k + a - 1$. The power of the q term then equals the power of the $q - \rho$ term in the Metropolis criterion (see Equation 7). At the end of the i^{th} cycle of 1000 generations, $i = 1, 2, \ldots, 10$, several summary values are recorded:

- A sample average of N: $\hat{E}_i^q[N] = \frac{1}{1000}\sum_{j=1}^{1000}\frac{k}{q_{ij}}$;
- A sample variance of N: $\widehat{Var}_i^q[N] = \frac{1}{999}\sum_{j=1}^{1000}\left(\frac{k}{q_{ij}} - \hat{E}_i^q[N]\right)^2$;
- An empirical distribution of N: $P_i^q[N = n] = \frac{1}{1000}\sum_{j=1}^{1000}\binom{n-1}{k-1}q_{ij}^k(1 - q_{ij})^{n-k}$ for a range of values of n.

(The superscript q indicates that the value was obtained by averaging across the corresponding

conditional values given the q_{ij}'s.) In addition, at the end of the 10 cycles, the following are calculated:

- An overall sample average of N: $\hat{E}^q[N] = \frac{1}{10000}\sum_{i=1}^{10}\sum_{j=1}^{1000}\frac{k}{q_{ij}} = \frac{1}{10}\sum_{i=1}^{10}\hat{E}_i^q[N]$;
- An overall sample variance: $\widehat{Var}^q[N] = \frac{1}{9}\sum_{i=1}^{10}\left(\hat{E}_i^q[N] - \hat{E}^q[N]\right)^2$.

The empirical distributions (frequency histograms) of N are first visually compared across the 10 cycles to note the stability of the iterations within each set of parameter values. The graphs are then visually compared across values of $\hat{Q}$ and ρ to note the variability and trends.

3.2 Simulation Results

Two aspects of the results are under consideration here:

i. Behavior of the sample expected values and standard errors across parameter values;

ii. Behavior of the empirical distributions of N within and across parameter values.

Table 2 shows sample expected values and sample standard errors for the nine combinations of $\hat{Q}$ and ρ when $k = 5$. Table 3 shows the same information for the eleven combinations chosen when $k = 20$. Within a value for $\hat{Q}$, we can see that the spread of the distribution, as measured by the sample standard errors, generally increases as ρ decreases. The trend is more consistent for $\hat{Q} = 1/2$ than for $\hat{Q} = 9/10$ in both tables. Intuitively, this trend is expected, since a smaller value for ρ indicates that fewer non-significant studies are being published. This leads to greater uncertainty in how many unseen studies may have been done, which leads to a distribution on the total number

of studies with a larger variance. The distribution on N is not bounded above, but is bounded below, so larger and larger values of N will have greater probabilities of occurring. The expected values will consequently show an increasing trend as well, as is true within every value of $\hat{Q}$ but one (Simulations #18 - 20). (One erratic iteration of the ten in Simulation #19 enabled much larger values of N to occur.) The trend of increasing standard errors is more consistent within Table 2 than within Table 3. The increase in k, or the values used for ρ, may have led to greater instability in the approximating distribution.

When $k = 5$, the empirical distributions of N are *very* stable across the ten iterations within each of the nine sets of parameter values. (These graphs are not included here.) Across the values for ρ, but within a value for $\hat{Q}$, the graphs gradually become wider and flatter as ρ decreases, as expected (see Figure 1 for an example). The change is gradual and gives the impression that the estimation is not extremely sensitive to the choice of ρ here. We can see here graphically the numerical trends evident in Table 2: the expected value of N increases slightly as ρ decreases, and the variance of the distribution increases slightly as ρ decreases. When $k = 20$, however, the stability decreases somewhat. Within a set of parameter values, the variability across the ten iterations is occasionally great. (These graphs are also not included.) Across these eleven sets, we see the same trends as when $k = 5$, but the changes in width and height are more dramatic, especially in Simulations #10 - 17 (see Figure 1 for an example). Within the simulations for $\hat{Q} = 1/10$ (#18 - 20), however, the changes in the graphs across the values for ρ are very minor. Most likely this is a result of the very narrow range of values for ρ that are possible (0.0 - 0.1) for these k and $\hat{Q}$ values. Outside of the the context of a particular meta-analysis, it is difficult to make more specific conclusions. See Eberly (1994) for more details, and Section 4 for an example.

4 Application: Lead Exposure and IQ in Children

Needleman and Gatsonis (1990) details several meta-analyses of studies relating childhood lead exposure to IQ level. The studies were chosen from the population of all studies on lead exposure and children's neurobehavioral development published since 1972, as found in MEDLINE, meeting programs, and dissertations. Each published study is required by the authors to contain the following in order to be included in a meta-analysis:

i. Use of a multiple regression analysis;

ii. A continuous IQ level as the response variable;

iii. Lead as a main effect in the regression;

iv. Control for non-lead covariates in the regression.

12 studies satisfied these criteria, of which 7 measured blood lead and 5 measured tooth lead. Studies for which all needed information is available give the data found in Table 4 (taken directly from Needleman and Gatsonis, 1990, Table 5). It must be noted that neither IQ levels nor lead levels were necessarily measured in the same way across all studies or even within the blood or tooth lead groups.

We assume a one-sided null hypothesis of a positive effect of lead on IQ, *i.e.*, $H_0 : \beta_{lead} \geq 0$, where β_{lead} denotes the regression coefficient. First, we carry out a simple meta-analysis (based on Rosenthal, 1978) to obtain an overall Z-value and p-value for the hypothesis of interest:

$$Z_{overall}^{Blood} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} = \frac{-3.86 - 1.67 + \cdots - 1.8}{\sqrt{7}} = -5.35,$$

which gives a one-sided p-value of essentially zero. Likewise,

$$Z_{overall}^{Tooth} = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} = \frac{-3 - 2.23 + \cdots - 1.17}{\sqrt{5}} = -3.42,$$

giving a one-sided p-value of 0.0003. We have to take the t-values as approximate Z-values here; the sample sizes are large enough that this seems reasonable. These p-values are strong indications that the null hypothesis is false, but we don't yet know how representative our sample of 12 studies is. We run simulations as described in Section 3 in order to make an assessment of the reliability of our meta-analytic results. From the last column in Table 4, six of the seven observed blood lead studies and two of the five tooth lead studies give significant results at $\alpha = 0.05$. Hence, the Gibbs sampler will be run first with $k = 7$ and $\hat{r} = 6/7$, and second with $k = 5$ and $\hat{r} = 2/5$. The program is run to produce 5,000 generated numbers of each of N and Q in total. By Equation 1, then, we can choose several values for ρ and calculate the corresponding values for $\hat{q}$. In order to capture any trend as the value of ρ changes, we will take $\rho = 1/10$, $1/2$, and $9/10$. The prior parameters on Q (a and b) are each taken to be 5, as before. The value for ℓ is adjusted (up or down, as necessary) from its initial setting at $k + a - 1$ to ensure that the Metropolis sampling "accepts" at least 75% of the generated "candidate" values. At the end of the 5,000 cycles, the following are calculated: $\hat{E}^q[N] = \frac{1}{5000}\sum_{i=1}^{5000}\frac{k}{q_i}$ and $\widehat{Var}^q[N] = \frac{1}{4999}\sum_{i=1}^{4999}\left(k/q_i - \hat{E}^q[N]\right)^2$. The results are shown in Table 5.

Given the assumptions made about the prior distribution on Q, these results tell us that there could be about (7 - 7 =) 0 to (11 - 7 =) 4 blood lead studies on this hypothesis which were unseen. The researcher must now make his or her best guess at an appropriate value for ρ. In the most optimistic case, $\rho = 9/10$; *most* non-significant and *all* significant studies are published. In this case, we expect no unseen studies, so our sample of published studies can be considered entirely trustworthy. In the least optimistic case, $\rho = 1/10$ and *most* non-significant studies are

not published, whereas all significant studies are. In this case, we could have four unseen studies. If all of them are *strongly* non-significant, or significant in the opposite direction, it is possible that our combined p-value could be overturned. However, Needleman and Gatsonis (1990, p.677) makes a very good point: "Given the expense of conducting human studies of lead exposure and the amount of attention directed to this question, is it unlikely that this number of negative studies have escaped notice." Clearly knowledge of the subject matter is needed to make a judgment on the probable value for ρ. For the tooth lead studies, there could be about (5 - 5 =) 0 to (11 - 5 =) 7 unseen studies. As above, in the most optimistic case, we expect no unseen studies, and our meta-analysis' results seems trustworthy. In the least optimistic case, there could be more non-significant studies out there than studies on hand. The meta-analysis could be giving us very biased results. Again, though, it seems unlikely that results with strong conclusions contrary to published conclusions would not have been noticed.

In cases where individual study Z-values are available, it may be helpful to compare the simulation results to the standard source augmentation method, Rosenthal's fail-safe number (Rosenthal, 1979). Rosenthal's number (FS) calculates the number of unseen studies *averaging null results* (*i.e.*, a p-value of zero or a Z-value of 0.5) needed to bring a significant overall p-value to a specified level. The fail-safe numbers are based on the same method of combining Z-values that was used above. Since those two Z-values are both significant, it makes sense to calculate Rosenthal's estimates and compare them with our simulation results:

$$FS_{Blood} = \left[\frac{\left(\sum_{i=1}^{k} Z_i\right)^2}{(1.645)^2} - k\right]^+ = \left[\frac{(-3.86-1.67+\cdots-1.8)^2}{(1.645)^2} - 7\right]^+ = 66.99$$

$$FS_{Tooth} = \left[\frac{\left(\sum_{i=1}^{k} Z_i\right)^2}{(1.645)^2} - k\right]^+ = \left[\frac{(-3-2.23+\cdots-1.17)^2}{(1.645)^2} - 5\right]^+ = 16.63.$$

Hence, 66 unseen studies giving null results are needed to overturn this combined p-value of zero from the blood lead studies, while 16 are needed to overturn the 0.0003 from the tooth lead studies. Since only 5 to 11 studies of any kind (significant or not, published or not, measuring blood or tooth lead levels) are expected to be out there on average, it seems highly improbable that there are enough unseen null studies to overturn the p-value, no matter the value of ρ. Taken in concert, our results and the fail-safe numbers offer reassurance that the meta-analyses are reliable. Rosenthal (1979) offers his own guidelines on what an "unlikely" number of unseen studies might be. He suggests that some fields may consider 100 or 500 unseen studies plausible, whereas other fields may deem only 10 or 20 unlikely. Rosenthal's recommendation is to consider $5k + 10$ the level at which the number of unseen studies becomes implausible. The $5k$ suggests that it is unlikely that there are more than 5 times as many studies filed away as there are on hand, while 10 sets the minimum number of studies at 15 when $k = 1$. In this example, the cutoffs would be 45 and 35 for the blood and tooth studies, respectively.

As a caution, the p-values calculated in Equation 4 above are based on what may or may not be a good estimate of the overall Z-values. One must always keep in mind that there are many other ways to calculate an overall p-value, ones that, for example, take sample sizes or sample variances into account (see Rosenthal, 1978). Some of those methods could give non-significant overall results, in which case any consideration of FS is nonintuitive. In addition, since this is a one-sided hypothesis testing situation, the researcher must consider the possibility of unpublished studies that are significant in the *opposite* direction. Rosenthal's estimates are a useful (and possibly reassuring) comparison to make when the data are available to calculate them. However, they are strictly ad hoc estimates and the statistical properties associated with them are not known; caution should be used in interpreting their results.

5 Conclusions

We have derived a method for approximating the total number of studies done on a particular hypothesis, given a selection probability (ρ), a distribution of the probability of publication (Q), and a meta-analysis of k available studies. The theory is complex only in that it must adapt to circumvent practical computational difficulties (*i.e.*, Metropolis simulation and Gibbs sampling). One drawback of this theory, of course, is that the prior distribution on Q must be specified. Very few researchers will be able to choose parameter values for the Beta prior distribution with any degree of assuredness. This distribution can take on almost any form for various values of a and b, and it is unclear how those choices influence the results of the simulations. Further investigation should be done regarding the effect of varying a and b on the stability of the simulations and on the precision of the approximations. Reassuringly, though, research in Bayesian statistics has shown that posterior distributions can be robust to the choice of the prior distributions. See, for example, Berger (1993). Another potential problem lies in the violations of assumptions. It is conceivable that the probability of publication is *not* constant across studies. In situations where a great deal of funding is allocated for large-scale nationwide clinical trials, for example, it is almost a certainty that these results will be published, significant or not. As a related issue, the biggest criticism of the fail-safe numbers is that they fail to distinguish between studies which are significant due to a large effect size, and studies which are significant due to a large sample size. Our method sidesteps this criticism by requiring that all studies considered are roughly of the same size. In practice, however, this assumption is not likely to be satisfied.

Given the simulation program, these methods are easy to implement and easy to interpret. An obstacle to using this method in a specific application is that a value (or possibly values) for ρ must be chosen. A researcher must have a good familiarity with both the publication process and

the activities of other researchers in his or her field to be able to give a reliable estimate. We recommend, therefore, that the simulations always be run for a range of values for ρ. Hopefully, from personal experience, this range can at least be limited to only a small portion of the interval $[0, 1]$. The application of this theory would be much improved if a method for estimating ρ is developed. Another disadvantage is that any application of this theory can only start from a count of the number of significant studies (*i.e.*, to calculate z/k) not from individual p-values nor Z-values. It seems there is a loss of information at some level here. Sample sizes and sample variances from the studies under consideration do not affect this procedure, when ideally it seems they should. The next step is perhaps to consider a model that depends not only on ρ, but also on other relevant covariates. Either ρ could be modeled deterministically, by choosing some function of the covariates, or a prior distribution for ρ could be chosen. A further and perhaps more realistic generalization of another aspect of the model would be to let $Q = \delta R + \rho(1 - R)$, so that not all significant studies are assumed published.

Using a range of values for ρ and the Gibbs/Metropolis procedure, a reasonable picture of the number of unseen studies can be formed for a specific meta-analysis application. Using Rosenthal's fail-safe estimates, we can calculate a second indication of the reliability of a significant overall p-value. Combining the two procedures, we can make the following statements: if Rosenthal's estimates give numbers which are too large to be likely under the distribution for N, then a meta-analysis' conclusions can be considered reliable. If Rosenthal's estimates give numbers which are small enough to be likely to occur, then the meta-analysis' results are clearly questionable. Alternatively, a researcher with a good knowledge of his or her field can make a judgment based on the simulated distributions of N without reference to Rosenthal's estimates.

6 References

Bayarri, M.J. and DeGroot, M. (1986). "Bayesian analysis of selection models," Technical Report 365, Dept. Statistics, Carnegie-Mellon University.

Bayarri, M.J. and DeGroot, M. (1991). "The analysis of published significant results," Technical Report 91-21, Dept. Statistics, Carnegie-Mellon University.

Begg, C.B. and Berlin, J.A. (1988). "Publication bias: a problem in interpreting medical data," *Journal of the Royal Statistical Society - Series A*, **151**:419-463.

Begg, C.B. and Mazumdar, M. (1994). "Operating characteristics of a rank correlation test for publication bias," *Biometrics*, **50**:1088-1101.

Berger, J. (1993). "An overview of robust Bayesian analysis," *Test*, **3**:5-124.

Berlin, J.A., Begg, C.B. and Louis, T.A. (1989). "An assessment of publication bias using a sample of published clinical trials," *Journal of the American Statistical Association*, **84**:381-392.

Cleary, R.J. (1996). "An Application of Gibbs Sampling to Estimation in Meta-Analysis: Accounting for Publication Bias," to appear in *Journal of Educational and Behavioral Statistics.*

Dear and Begg (1992). "An approach for assessing publication bias prior to performing a meta-analysis," *Statistical Science*, **7**:237-245.

Dickersin, K., Min, Y.-I., and Meinert, C.L. (1992). "Factors influencing publication of research results," *Journal of the American Medical Association*, **267**(3):374-378.

Easterbrook, P.J., Berlin, J.A., Gopalan, R., and Matthews, D.R. (1991). "Publication bias in clinical research," *Lancet*, **337**:867-872.

Eberly, L.E. (1994). *Estimating the Number of Unseen Studies in a Meta-Analysis*, M.S. Thesis, Biometrics Unit, Cornell University, Ithaca, NY.

Frongillo, E. (1991). *Combining Information Using Hierarchical Models*, Ph.D. Dissertation, Biometrics Unit, Cornell University, Ithaca, NY.

Gelfand, A.E. and Smith, A.F. (1990). "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, **85**:398-409.

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**:721-741.

Glass, G.V., McGaw, B., and Smith, M.L. (1981). *Meta-Analysis in Social Research*, Beverly Hills, CA: Sage Publications.

Hedges, L.V. (1984). "Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences," *Journal of Educational Statistics*, **9**:61-85.

Hedges, L.V. (1992). "Modeling publication selection effects in meta-analysis," *Statistical Science*, **7**:246-255.

Iyengar, S. and Greenhouse, J.B. (1988). "Selection models and the file drawer problem," *Statistical Science*, **3**:109-135.

Light, R. and Pillemer, D. (1984). *Summing Up: the Science of Reviewing Research*, Cambridge, MA: Harvard University Press.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, **21**:1087-1091.

Needleman, H.L. and Gatsonis, C.A. (1990). "Low-level lead exposure and the IQ of children," *Journal of the American Medical Association*, **263**:673-678.

Robert, C.P. (1990). *The Bayesian Choice: A Decision-Theoretic Motivation*, New York: Springer-Verlag.

Rosenthal, R. (1978). "Combining results of independent studies," *Psychological Bulletin*, **85**:185-193.

Rosenthal, R. (1979). "The "file drawer" problem and tolerance for null results," *Psychological Bulletin*, **86**:638-641.

Tanner, M.A. and Wong W.H. (1987). "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, **82**:528-540.

White, K.R. (1982). "The relation between socioeconomic status and academic achievement," *Psychological Bulletin*, **91**:461-481.

7 Tables and Figures

Table 1: Parameter values by Simulation

$k = 5$				$k = 20$			
Simulation	$\hat{Q}$	ρ	$r = z/k$	Simulation	$\hat{Q}$	ρ	$r = z/k$
1	0.9	0.900	0.0	10	0.9	0.900	0.00
2		0.875	0.2	11		0.867	0.25
3		0.833	0.4	12		0.800	0.50
4		0.750	0.6	13		0.600	0.75
5		0.500	0.8	14		0.000	0.90
6	0.5	0.500	0.0	15	0.5	0.500	0.00
7		0.375	0.2	16		0.333	0.25
8		0.167	0.4	17		0.000	0.50
9	0.1	0.100	0.0	18	0.1	0.100	0.00
				19		0.053	0.05
				20		0.000	0.10

Table 2: Expected values and standard errors when $k = 5$

Simulation	$\hat{Q}$	ρ	$\hat{E}^q[N]$	$(\widehat{SE}^q[N])$
1	0.9	0.900	5.43	(0.04)
2		0.875	5.49	(0.04)
3		0.833	5.60	(0.04)
4		0.750	5.85	(0.08)
5		0.500	6.64	(0.13)
6	0.5	0.500	8.17	(0.26)
7		0.375	9.04	(0.39)
8		0.167	10.73	(1.47)
9	0.1	0.100	16.72	(3.00)

Table 3: Expected values and standard errors when $k = 20$

Simulation	$\hat{Q}$	ρ	$\hat{E}^q[N]$	$(\widehat{SE}^q[N])$
10	0.9	0.900	21.50	(0.06)
11		0.867	21.73	(0.06)
12		0.800	22.18	(0.20)
13		0.600	23.24	(0.55)
14		0.000	26.45	(0.19)
15	0.5	0.500	32.08	(0.65)
16		0.333	37.41	(1.60)
17		0.000	41.35	(0.70)
18	0.1	0.100	85.42	(4.82)
19		0.053	103.00	(19.67)
20		0.000	96.88	(2.78)

Figure 1: Simulated distributions of N

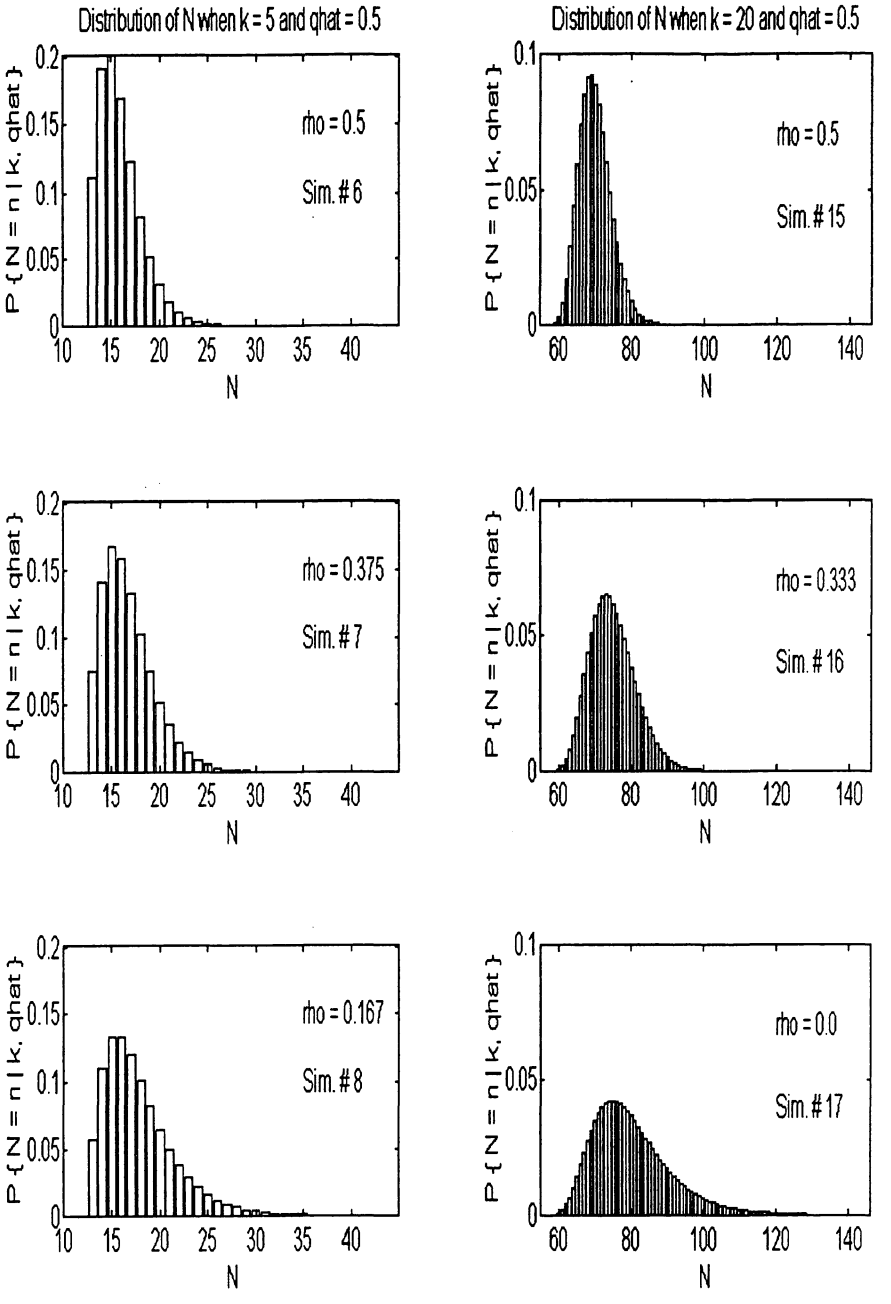

Table 4: Lead Coefficients for Full-scale IQ Scores

Study	Regression coefficient	Standard error	t-value	Sample size	One-sided p-value
		Blood Lead Studies			
Hatzakis, *et al.*	-0.27	0.07[a]	-3.86[a]	509	0.0001
Hawk, *et al.*	-0.25	0.15	-1.67	75	0.05
Schroeder, *et al.*	-0.2	0.07[a]	-2.78	104	0.003
Fulton, *et al.*[b]	-3.7	1.37	-2.77	501	0.003
Yule, *et al.*[b]	-8.08	4.63	-1.75	129	0.04
Lansdown, *et al.*[b]	2.15	4.48[a]	0.48	86	0.68
Emhart, *et al.*	NA[c]	NA	-1.8[a]	80	0.04
		Tooth Lead Studies			
Needleman, *et al.*	-0.21	0.07	-3	218	0.001
Hansen, *et al.*	-4.27	1.91	-2.23[d]	156	0.01
Winneke, *et al.*	-0.13	4.66	-0.03[d]	115	0.49
Pocock, *et al.*[b]	-0.77	0.63	-1.22	388	0.11
Fergusson, *et al.*[b]	-1.46	1.25	-1.17	724	0.12

[a]Estimated from data in article.
[b]Used log transformation.
[c]Not available.
[d]Obtained from the author.

Table 5: Expected values and standard errors

Simulation	ρ	$\hat{q}$	$\hat{E}^q[N]$	$(\widehat{SE}^q[N])$
Blood Lead Studies: $k = 7$ *and* $\hat{r} = 6/7$				
1	0.1	0.87	10.69	(1.87)
2	0.5	0.93	8.94	(0)
3	0.9	0.99	7.30	(0)
Tooth Lead Studies: $k = 5$ *and* $\hat{r} = 2/5$				
4	0.1	0.46	10.97	(0.01)
5	0.5	0.70	8.03	(0)
6	0.9	0.94	5.43	(0)